Available online at www.sciencedirect.com
SciVerse ScienceDirect
Procedia Engineering 29 (2012) 981 – 985
Procedia Engineering
www.elsevier.com/locate/procedia

2012 International Workshop on Information and Electronics Engineering (IWIEE)

Design of Automatic Question Answering System Base on CBR

LIANG Zhenqiu [a*]

Guangdong Polytechnic Normal University,No. 293 Zhongshan Road,Guangzhou,510665, China

Abstract

Against the lack of existing answering system in terms of intelligence and human-computer interaction, on the basis of the CBR, proposed a design of intelligent question answering system, and study the key technologies of this program, include the automatic segmentation , questions similarity calculation, improve search efficiency. Experiments show that the program can improve the accuracy and intelligence of answering system, has some practical value.

Keyword: case based reasoning (CBR)； Intelligent question answering system；Automatic Word Segmentation； Question similarity

1. Introduction

With the continuous development of distance education, answering system has become an important component of the online education platform. At present, answering system is gradually developed from artificial answering to intelligence machine answering automatically. Because human beings to solve problems, often based on past similar approach, after modified appropriately to solve new problems. Based on this idea, this paper presents an answering system based on case-based reasoning; its main idea is to automatically retrieve existing and valid historical cases to answer the new questions. This article describes the CBR-based automated answering system working principle, system architecture and

* Corresponding author. Tel.: +86-18924022292.
E-mail address: LiangLaoShi@126.com.

doi:10.1016/j.proeng.2012.01.075

working process, automatic segmentation, similarity calculation questions, improve search efficiency and other key technologies are described in detail.

2. The principle of CBR applied to intelligent question answering system

CBR is a problem-solving technology based on experience, and has been widely used in many fields. It is use the earlier examples which solve a similar problem and stored in the library, and to explain its reasoning by reference examples. CBR has the advantage of easy access to knowledge that compared with traditional rule-based reasoning solution method, and avoid the bottleneck problem when traditional knowledge systems getting knowledge. It has advantages such as easy maintenance the knowledge database and do not need domain experts interference, etc. CBR applied to the answering system, classify and extract the questions and answers which stored in library, build a case base that data accurate, complete, organized, easy retrieval and maintenance, can greatly improve the level of intelligent question answering system.

3. Intelligent answering system structure based on CBR

By analyzing the basic operation process of the answering system, the system workflow as follow:

1) Student enters the problem into answering system;
2) System parse the question, extract the key words by thesaurus-based mechanical segmentation;
3) Search the historical questions database according the keywords to get the candidate questions.
4) Calculate the similarity between the new questions and candidate history questions
5) If the calculated similarity of historical question achieved the required threshold, then directly go to steps 6 to show the answer. Otherwise, go to step 8 to enter the full-text search module.
6) Shows the answers of retrieved history questions.
7) Updated answering database, if the difference between students questions and history questions exceeds a certain threshold (for example 20%), then the student questions as a new question, recorded in the answering database, and then go to steps 10.
8) Full-text search module, according the keywords of student questions to search full-text.
9) If the calculated similarity achieves the required threshold, then display results of full-text search, otherwise record this question and wait for teacher to answer.
10) End of the program.

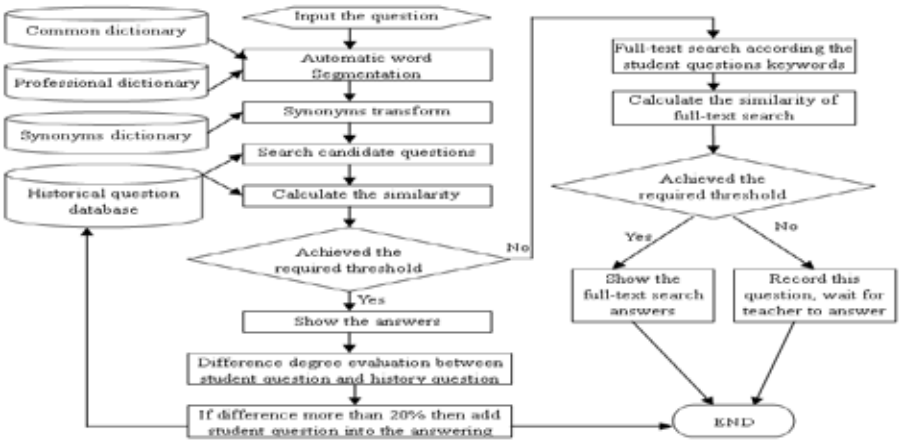

Fig. 1. The working process of intelligent question answering system

4. Intelligent answering system structure based on CBR

4.1. Automatic word segmentation

The establishment of word segmentation dictionary and choice word segmentation algorithm is the most important step of dictionary-based automatic word segmentation.

4.1.1. The establishment of word segmentation dictionary

The word segmentation dictionary includes common dictionary, professional dictionary and synonyms dictionary. The common dictionary contains some common words that may be used in the new problems; the common thesaurus in this system uses a txt version of Peking University download from network. In addition, some common interrogative need to add into the commonly used word dictionary, such as "why", "what", "how", etc. Professional dictionary is established based on the specific subjects, including the professional words of corresponding subject, with good specificity and discrepancy of digestion. Because this answering system faces the computer-related courses, so the professional vocabulary uses the download "English-Chinese computer dictionary". Synonym dictionary uses the txt version download from network. Through the cutting and processing to remove duplicate fields, the thesaurus of SQL Server version generated finally.

4.1.2. The word segmentation algorithm

There are many Chinese word segmentation system based on mechanical segmentation algorithm, and their word segmentation correct rate is also high, so this system using the maximum matching algorithm to complete the word segmentation of questions. The main steps of maximum matching algorithm are: Assume the longest word in automatic segmentation dictionary is N-words, then take the first N words of current question as the matching field and compare with the thesaurus. If match that show this N chars is a word, then cut out this word from the question and recorded. If can not matched, then reduce a word and continue match in the dictionary. Until only remaining one word, the algorithm terminates. If this string cannot be split, then record it as a new word. Algorithm flow is shown below.

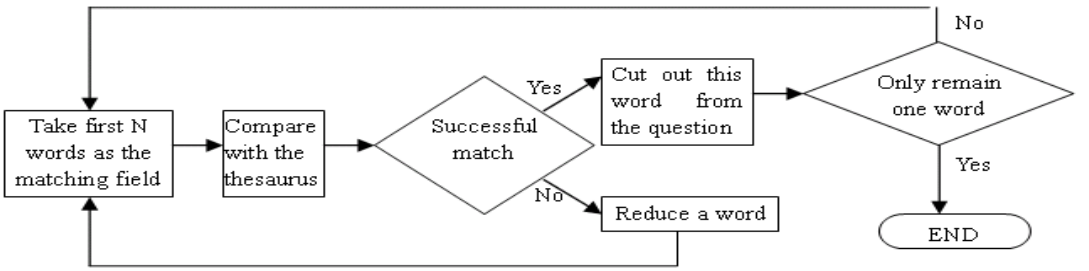

Fig. 1. The maximum matching algorithm flowchart

4.2. Similarity calculation of questions

In answering system, usually the questions that raised by students are relative short. In most cases the questions is only one sentence. So you can calculate the similarity of student's questions and the history question that in answering database after completed word segmentation. Because of the same question may be have different expression, so can not just compare the number of the same keyword, but also considering the number of keywords between two sentence and so on. After a comprehensive calculation, if the calculated similarity achieved the required threshold, then we can consider that the history question

is the same problem with the user questions corresponding, can display the answer. Otherwise, we can believe that the answering database without the questions which asked by users, so go to the full-text search module.

Assume that there are user question X and standard question Y, K(X) denotes the keyword number of X, K(Y) denotes the keyword number of Y, S(X,Y) denotes the same keyword number between X and Y. The questions similarity Like (X, Y) between X and Y, can be calculated by formula (1).

$$Like(X,Y)= \lambda 1*KeyLike(X，Y)+\lambda 2*KeyLong(X，Y) \quad (1)$$

Thereinto, KeyLike (X, Y) is the proportion of the same words which between user question X and standard question Y. The larger the ratio, the higher the similarity, it can be calculated by formula (2); KeyLong (X, Y) is the proportional relationship of the number of keywords between sentence, the maximum value is 1, indicates the keywords number between two sentence is the same, it can be calculated by formula (3).

$$KeyLike\ (X,Y) =2*S(X，Y) / (K(X)+K(Y)) \quad (2)$$

$$KeyLong\ (X,Y) =1-abs(K(X)-K(Y)) / (K(X)+K(Y)) \quad (3)$$

λ1 and λ2 is the adjustment coefficient, regulate the scale factor of the same keyword and the proportion of keyword number coefficient, and want to satisfy λ1 + λ2 = 1, for example, λ1 = 0.7, λ2 = 0.3. Finally, the greater of questions similarity calculated, the more similar of user questions X and the standard questions. 1 means that X and Y are identical, 0 indicates that they are completely different.

4.3. Answering database updates and efficiency gains

The process of using the system will continue emergence some new questions and answers. Since only continuous accumulation the new problems, the intelligent of answering system to be continuous improvement, so the system will automatically add new questions and answers into the answering database.

Meanwhile, to improve the speed of retrieval and comparison, answering system use the method of redundant storage the keyword to database. That is after keywords segmentation of each questions, store it in another field. If the keyword has synonyms, then replaced the keyword by its synonym, and save it into the answering database. When calculate similarity between student's questions and history questions, only need to carry keyword segmentation through for student's questions and find synonyms for it. The sentences of answering database is no longer need to carry keyword segmentation operation, but also no longer need to search synonymous, just load the keywords from database which had finished keyword segmentation and had replaced by synonymous. This is a method that expense storage space to save computation time, in the present storage capacity of the current computer, waste the storage space in redundant storage is almost negligible. Using this storage method can effectively improve the system response speed.

Answering database have eight fields: QuestionID: the serial No of questions; Question: the text of question; QuestionType: the type of questions; Keywords: the keywords from word segmentation operation; Same_keyword1: the keyword sequence 1 that search from thesaurus and replaced by the synonymous; Same_keyword2: the keyword sequence 2 that search from thesaurus and replaced by the synonymous; Same_keyword3: the keyword sequence 3 that search from thesaurus and replaced by the synonymous; Answer: the text of correct answer.

When search in answering database, the first, search for the history questions by the word segmentation keywords, determine the set of candidate historical questions. And then, calculate the similarity between

the new questions and the candidate history questions, the task boils down to determine the similarity between sentences.

5. Conclusion

Automated answering system is the important components of distance teaching platform, its development and consummate depend on various comprehensive technologies. Use CBR technology to build the answering database, is an effective way for improve the intelligent of answering system. This paper introduces the automatic question answering system based on CBR. This system can analyze the questions that inputted in natural language, search for candidate question set in historical question database by the keyword automatically. Through calculate the similarity of sentences, the similar historical answers are returned to the user. The practice shows that the system already achieved practical results in answering accuracy and intelligent, in distance education and other related fields has a certain practical value.

References

[1] ZHENG Geng-zhong. The study and implement of case base in intelligent answering sys tem based on CBR. [J]. Microcomputer Information, 2008(24):259-261.

[2] GUO Xiao-yan, ZHANG Bo-feng, FANG Ai-guo.Research on question similarity algorithm for intelligent question answering system and its implementation [J]. Computer Applications. 2005(02):449-452

[3] KANG Wen-ning, YANG Zhi-iang. Research and Application of Sentence Similarity Measurement in Intelligent Answering System [J]. Computer Technology and Development. 2010(02):71-74

[4] WU Daiwen, YANG Fangqi. The Performance Study of Database Full-Text Retrieval Based on Lucene [J] Microcomputer Applications. 2011(06):53-58

[5] ZHOU Fa-guo, YANG Bing-ru. New method for sentence similarity computing and its application in question answering system [J]. Computer Engineering and Applications. 2008(01):165-178

[6] QIN Bing,etc.Question answering system based on frequently asked questions[J]. Journal of Harbin Institute of Technology. 2003(10):1179-1182